PROYECTO FINAL- ENTREGA 3

DAVID MEJÍA CASTAÑO, 1007221842
DIEGO HERNANDO ARANGO RÍOS, 1042768156

INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL PARA LAS CIENCIAS E INGENIERÍAS / MODELOS Y SIMULACIÓN DE SISTEMAS I

UNIVERSIDAD DE ANTIOQUIA

DOCENTE
RAÚL RAMOS POLLÁN

INTRODUCCION

Para el avance de la segunda entrega hemos explorados unos datos categóricos y numéricos para correlación de variables teniendo en cuenta que tan dependientes o independientes hay entre ellas. Luego seleccionar la que mejor precisión se acerca para la construcción de nuestro modelo.

La exploración de datos y el análisis estadístico son elementos fundamentales para comprender la complejidad de conjuntos de datos. En este contexto, se llevó a cabo un exhaustivo examen de variables, destacando las correlaciones mediante una matriz. En particular, se observaron relaciones significativas entre variables numéricas, centrándose especialmente en las unidades curriculares de distintos semestres. Este análisis se complementó con representaciones gráficas que permitieron visualizar las tendencias y patrones de correlación.

Una vez identificadas estas correlaciones, se procedió a analizarlas en función de la variable objetivo, desglosando la información para entender cómo se distribuyen las relaciones en diferentes categorías. Este enfoque proporciona una visión más detallada de cómo las variables se relacionan entre sí y con la variable objetivo.

Posteriormente, se emprendió la tarea de entrenar y evaluar diversos modelos, desde métodos más simples hasta implementaciones de redes neuronales. Se abordó la división del conjunto de datos y se examinaron modelos como Regresión Logística, Árboles de Decisión, Bosques Aleatorios y Redes Neuronales. Cada modelo se evaluó en términos de precisión, revelando fortalezas y limitaciones.

Este proceso de exploración y modelado ofrece una panorámica integral para comprender la complejidad de los datos y, en última instancia, identificar patrones predictivos que permitan tomar decisiones informadas. En este contexto, se destaca la eficacia de la red neuronal como un enfoque prometedor para abordar los desafíos asociados con la escasez de datos y las relaciones no lineales en conjuntos de datos complejos.

Exploración de datos:

Para la exploración de los datos se realizó inicialmente un análisis general de las variables, observando a partir de una matriz de correlación las correlaciones existentes entre las diferentes variables numéricas:

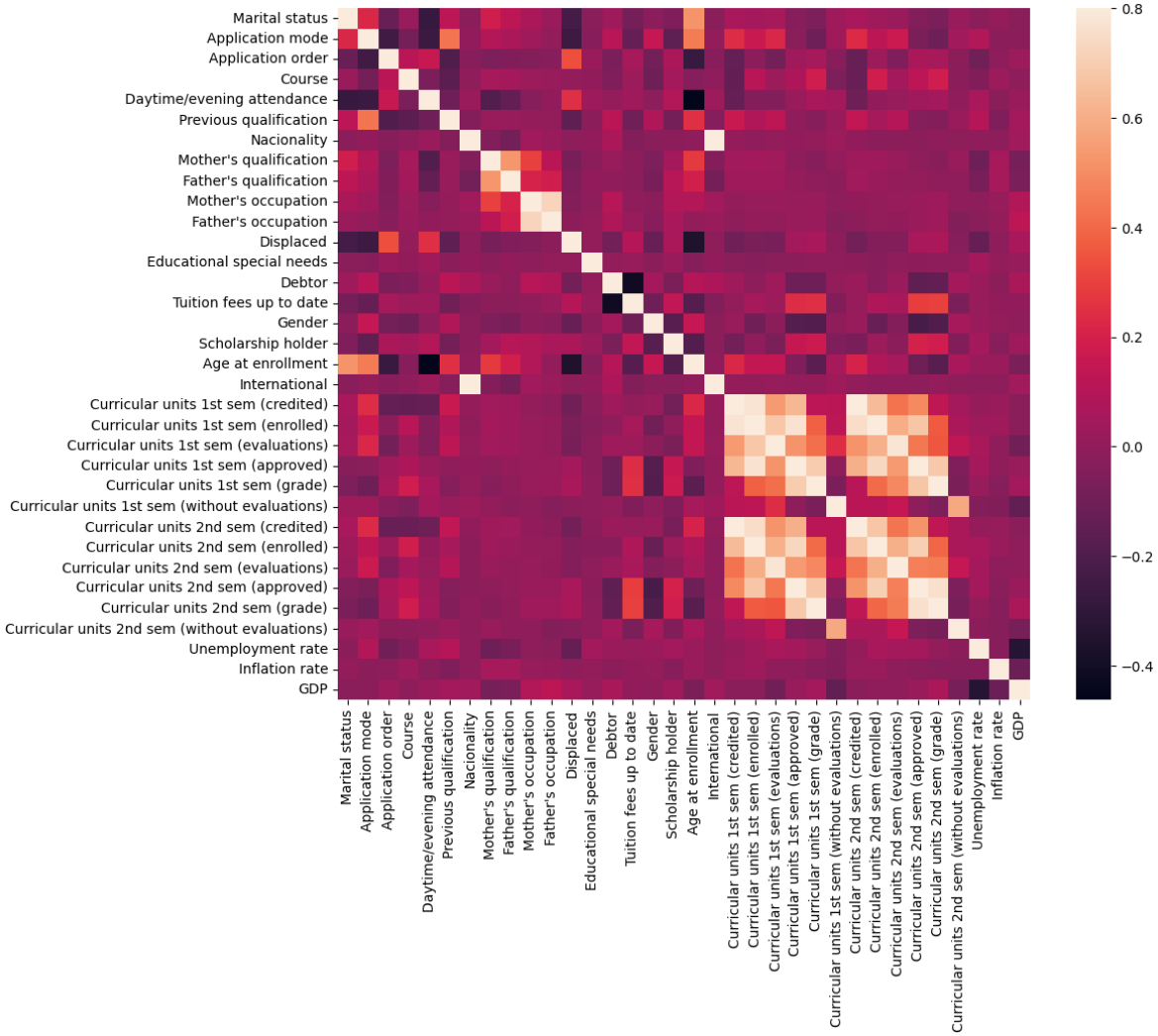

Figura 1: Matriz de correlación entre variables

Las correlaciones más fuertes halladas entre las variables para valores positivos se encuentran entre **International** y **Nacionality** y otra muy interesante entre las variables correspondientes a Curricular entre 1sem y 2 sem en las mismas categorías (**credited ,enrolled ,evaluations ,approved ,grade ,without evaluations**) y una correlación pero un poco más baja entre las diferentes categorías, esto se puede observar fácilmente en la figura anterior en los cuatro cuadros que resaltan por su color claro en la parte inferior derecha. Por otro lado, las variables **Age at enrollment** se encuentra correlacionada negativamente con **Daytime/evening attendance** y **displaced**, además de las variables **Tuition fees up to date** con **Debtor**.

Al analizar a partir de gráficos las correlaciones identificadas entre las variables a excepción de las correspondiente a Curricular se obtiene la siguiente gráfica:

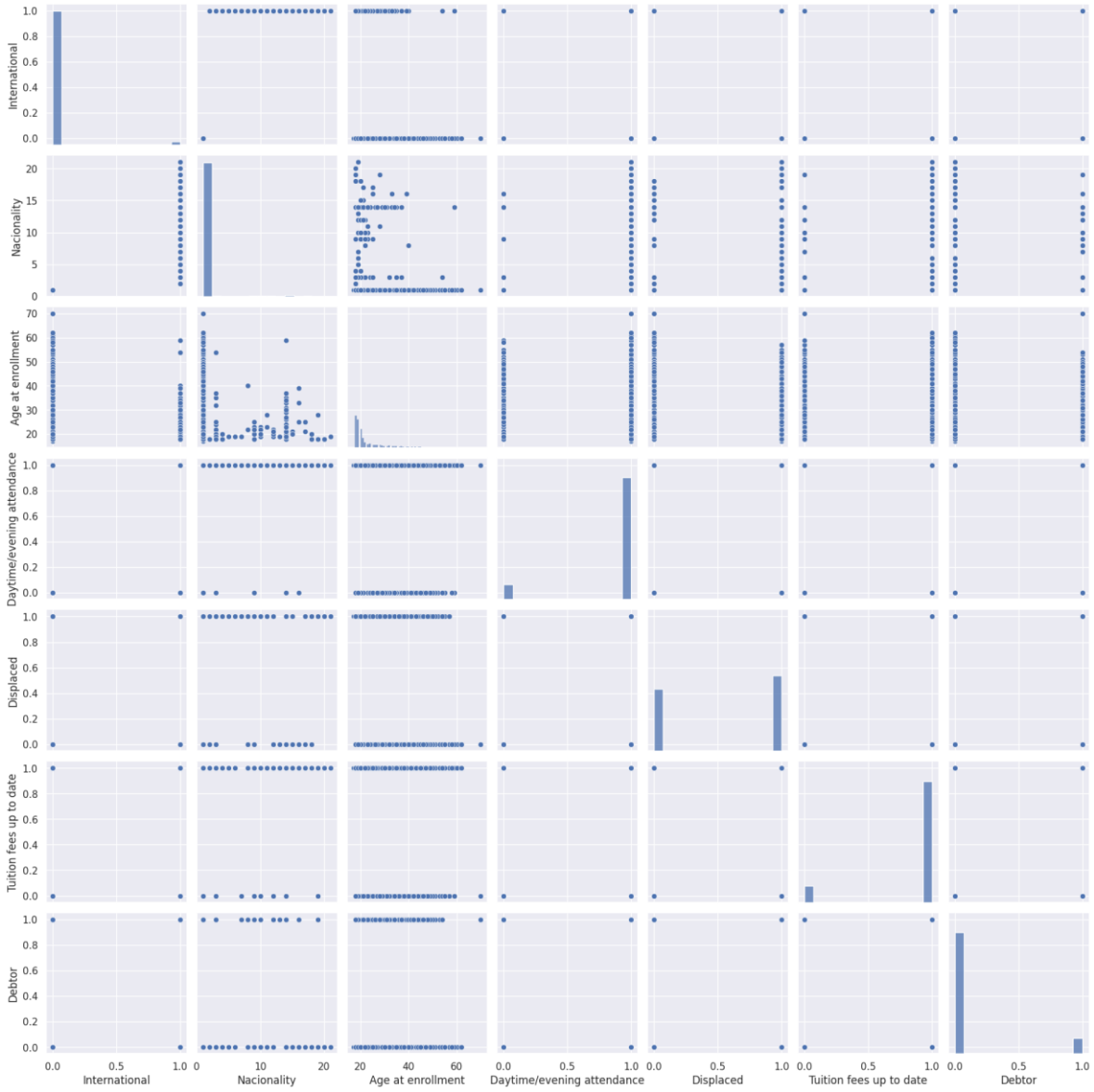

Figura 2: Análisis gráfico correlación entre variables

Las cuales aun que denotan cierta relación entre estas variables, al ser categóricas y con pocas categorías no nos permite establecer una correlación real que sea relevante para nuestro estudio.

Por esto procedemos a analizar las correlaciones existentes entre las variables correspondientes a Curricular, observando claras relaciones lineales entre las diferentes variables, como se observa claramente en los siguientes gráficos:

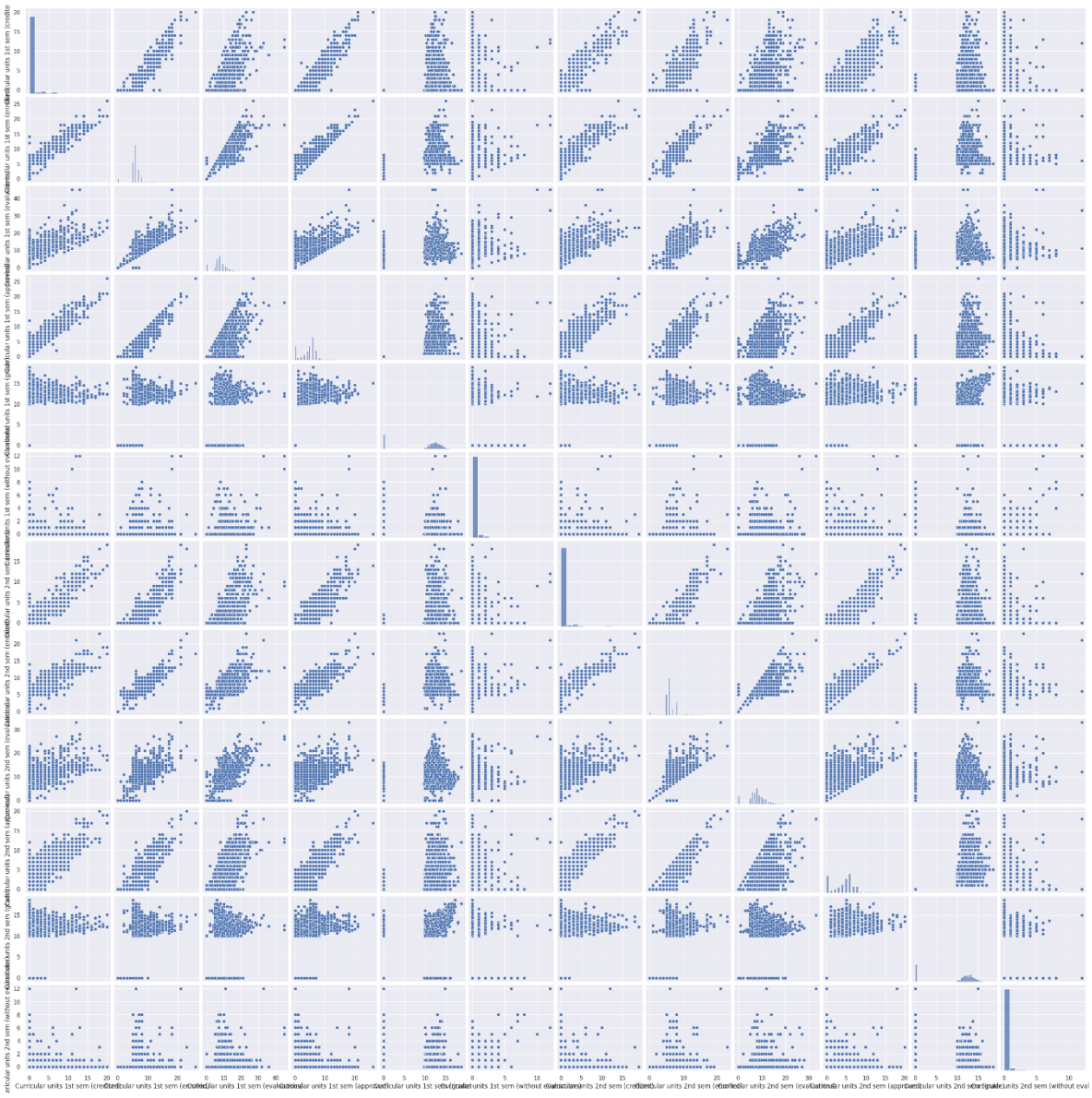

Figura 3: Análisis gráfico correlación entre variables Curricular

Permitiéndonos concluir que las siguientes variables contienen una correlación positiva:

- Curricular units 1st sem (credited) con Curricular units 1st sem (enrolled)

- Curricular units 1st sem (credited) con Curricular units 1st sem (approved)

- Curricular units 1st sem (credited) con Curricular units 2nd sem (credited)

- Curricular units 1st sem (credited) con Curricular units 2nd sem (enrolled)

- Curricular units 1st sem (credited) con Curricular units 2nd sem (approved)

- Curricular units 1st sem (enrolled) con Curricular units 1st sem (evaluations)

- Curricular units 1st sem (enrolled) con Curricular units 1st sem (approved)

- Curricular units 1st sem (enrolled) con Curricular units 2nd sem (credited)

- Curricular units 1st sem (enrolled) con Curricular units 2nd sem (enrolled)

- Curricular units 1st sem (enrolled) con Curricular units 2nd sem (approved)

- Curricular units 1st sem (evaluations) con Curricular units 1st sem (approved)

- Curricular units 1st sem (evaluations) con Curricular units 2nd sem (enrolled)

- Curricular units 1st sem (evaluations) con Curricular units 2nd sem (approved)

- Curricular units 1st sem (approved) con Curricular units 2nd sem (credited)

- Curricular units 1st sem (approved) con Curricular units 2nd sem (enrolled)

- Curricular units 1st sem (approved) con Curricular units 2nd sem (approved)

- Curricular units 2nd sem (credited) con Curricular units 2nd sem (enrolled)

- Curricular units 2nd sem (credited) con Curricular units 2nd sem (approved)

- Curricular units 2nd sem (enrolled) con Curricular units 2nd sem (evaluations)

- Curricular units 2nd sem (enrolled) con Curricular units 2nd sem (approved)

Tal como se observa en la siguiente matriz de correlación enfocada solo en las características con mayor correlación encontrada:

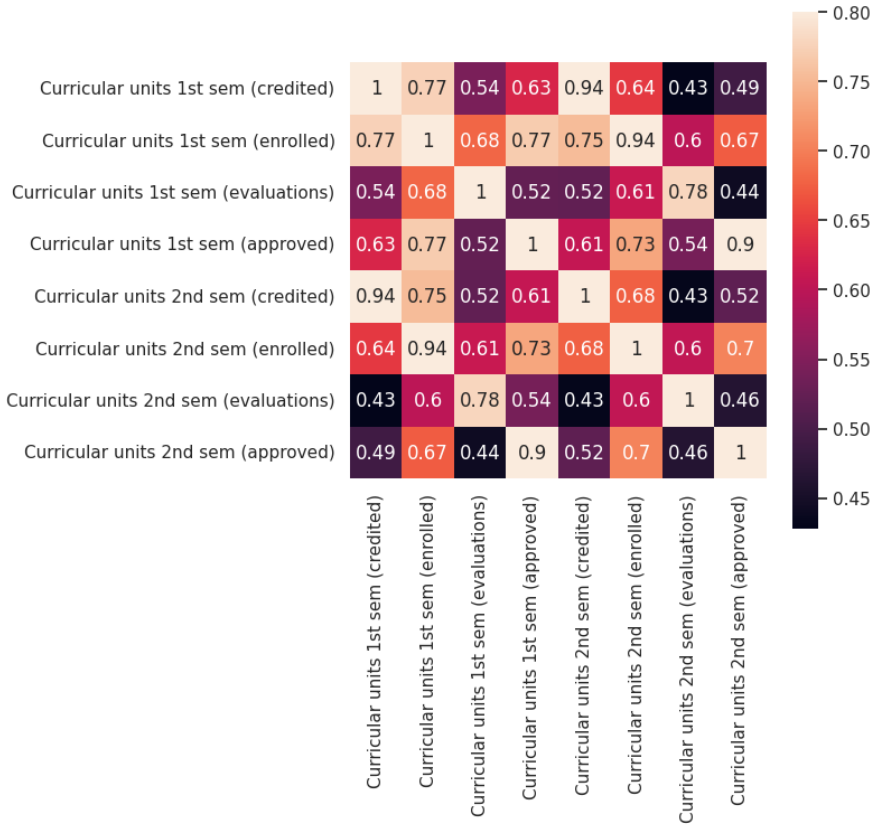

Figura 4: Matriz de correlación entre variables de Curricular con correlación alta.

Donde se observa que la mayor correlación cercana al 90% para todas las características se encuentra en las variables correspondientes a la misma característica en el primer y segundo semestre. De lo cual concluimos que solo es necesario tener en cuenta las variables curricular para el primer semestre ya que las de segundo semestre no aportan información a excepción de las correspondientes a **grade** y **without evaluations.**

Finalmente analizando la variable target, observamos que son mayor la cantidad de datos correspondientes a graduados y de abandono que de matriculados.

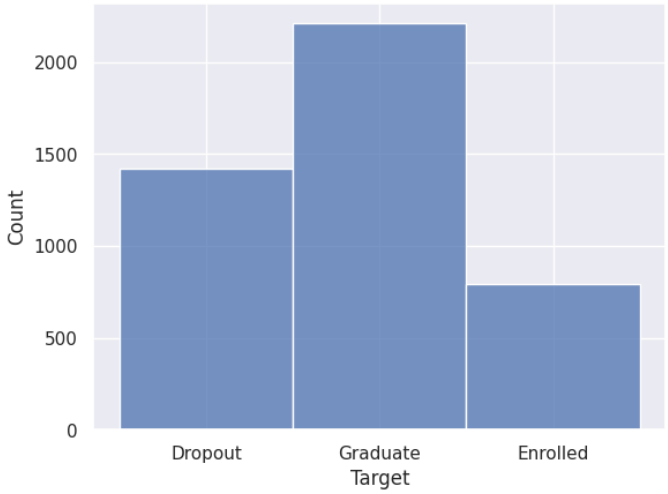

Figura 5: Histograma variable Target

Por otro lado, al analizar las correlaciones mencionadas anteriormente y diferenciándolas por target, se aprecia en algunos datos cierta segmentación dependiente al target, sin embargo, no hay ninguna relación de variables que brinde una segmentación suficientemente fuerte para enfocarnos solo en esto (Figura 6).

Entrenamiento y evaluación de modelos implementados:

Para esta fase de entrenamiento de modelos, se propusieron diferentes modelos desde los más simples hasta los más complejos, los modelos propuestos son:

- LogisticRegression
- DecisionTreeClassifier
- RandomForestClassifier
- Red Neuronal

Para el entrenamiento de estos modelos se dividio el dataset en dos subconjuntos, uno de entrenamiento con el 80% de los datos y otro de testeo con el 20% de estos, para esto se hizo uso de la función ***train_test_split*** de sklearn, que reparte los datos de forma aleatoria y bajo la misma proporción correspondiente a cada target como se observa en la matriz (Figura 7).

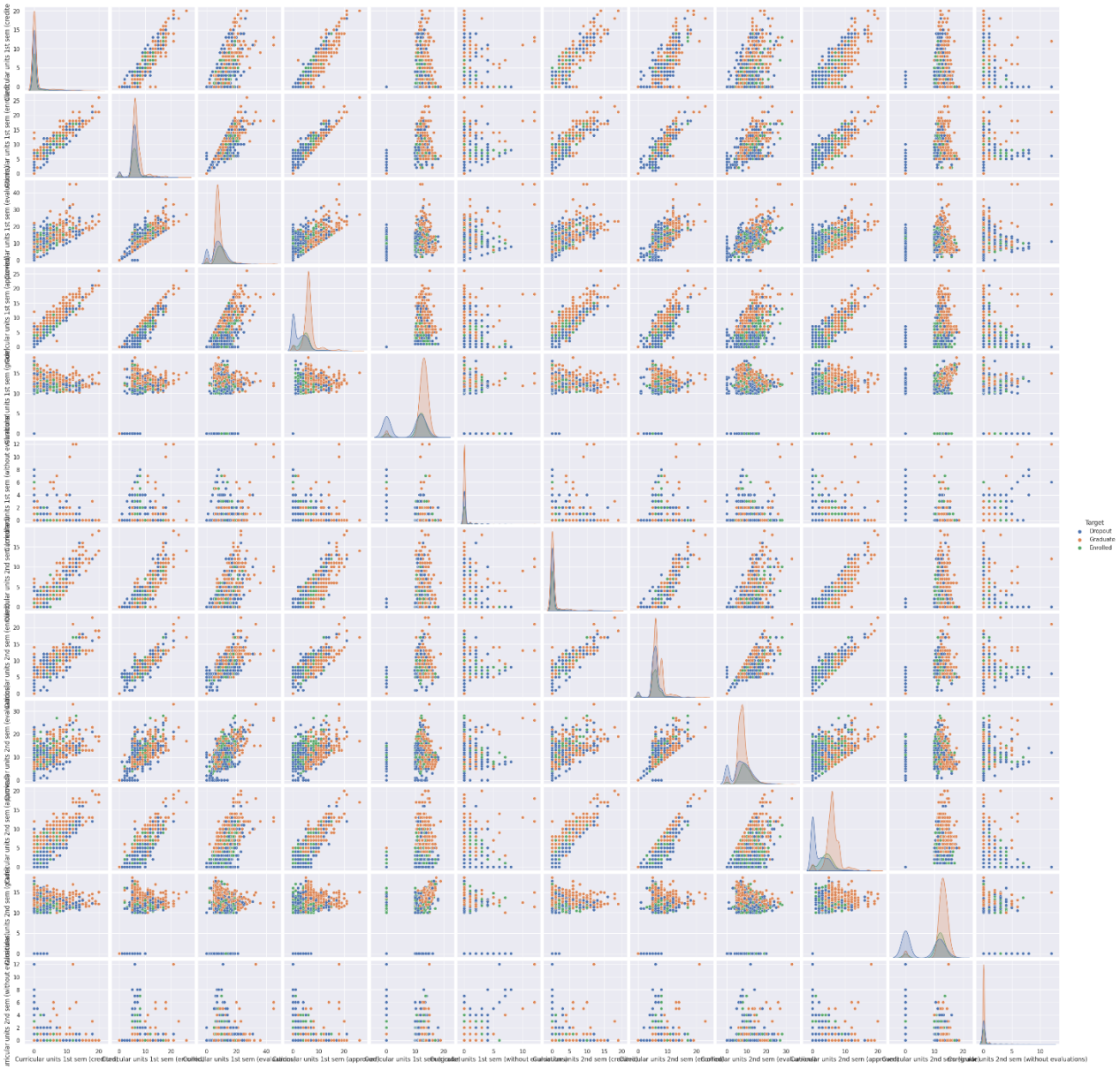

Figura 6: Análisis gráfico de correlaciones entre variables Curricular segmentado por Target

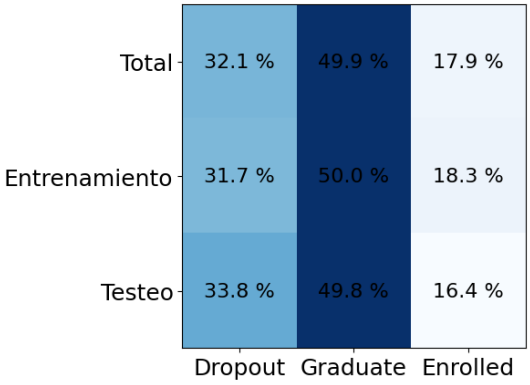

Figura 7: Matriz de porcentajes de targets en cada conjunto.

LogisticRegression:

Se entreno este modelo a partir de la librería sklearn.linear_model de la cual se obtuvo el modelo, obteniendo un accuracy de 78.24% para el conjunto de entrenamiento y de 74.24% para el de testeo y una matriz de confusión en la predicción del conjunto de testeo (Figura 8), la cual denota un buen comportamiento en la predicción del Target correspondiente a **Dropout** y **Graduate**, pero un bajo rendimiento en la predicción de **Enrolled**. Esto puede deberse a la baja cantidad de datos correspondiente a Enrolled que se poseen en comparación a las otras categorías.

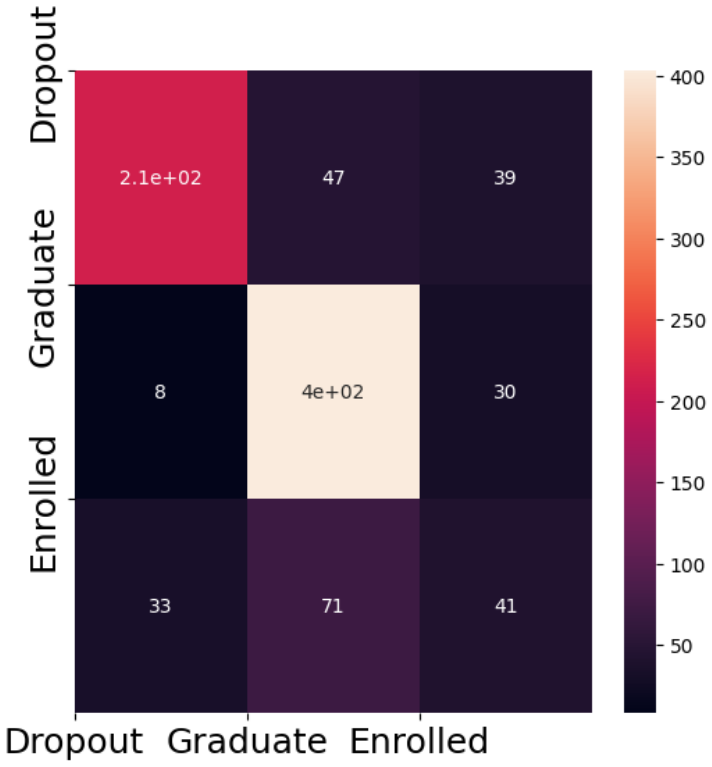

Figura 8: Matriz de confusión evaluación del modelo en conjunto de testeo

DecisionTreeClassifier:

Igual al anterior se entreno este modelo el cual se obtuvo a partir de la librería sklearn.tree obteniendo un accuracy de 100% para el conjunto de entrenamiento y de 67.23% para el conjunto de testeo, evidenciando un claro sobre ajuste ya que predice perfectamente el conjunto de entrenamiento, pero falla con el de testeo tal como se observa en la matriz de confusión (Figura 9).

Por esto se opta por usar cross_validation, buscando encontrar el modelo entrenado con el conjunto adecuado que nos permita mejorar este comportamiento. Al aplicar esto con 10 divisiones se encuentra un accuracy para el conjunto de testeo promedio de 74.5 ±1.5 % y el de entrenamiento de 76.9±0.6 % denotando que la mejora no es mucha para el conjunto de testeo del accuracy obtenido.

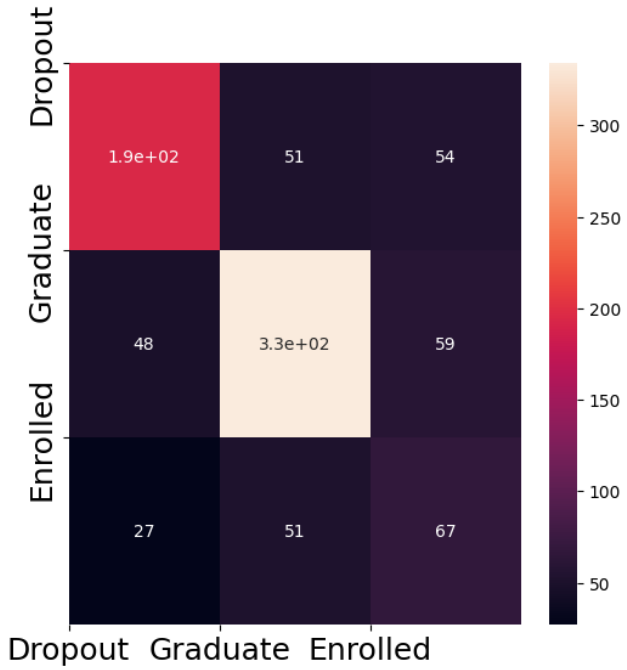

Figura 9: Matriz de confusión conjunto de testeo evaluado en modelo **DecisionTreeClassifier**

RandomForestClassifier:

Análogo al modelo anterior, se aplica el modelo de **RandomForestClassifier** haciendo uso de la librería sklearn.ensemble con cross validation encontrando un accuracy promedio para el conjunto de testeo de 75 ± 1 % y de entrenamiento de 76.6 5 ±0.5 % para 10 divisiones, permitiéndonos encontrar el mejor modelo con un accuracy de 100% en entrenamiento y 75.25 % en testeo lo cual mejora un poco los resultados obtenidos anteriormente, tal como se observa en su correspondiente matriz de confusión (Figura 10), aun que en esta se observa un deterioro para la variable Enrrolled la predicción de las otras variables mejora.

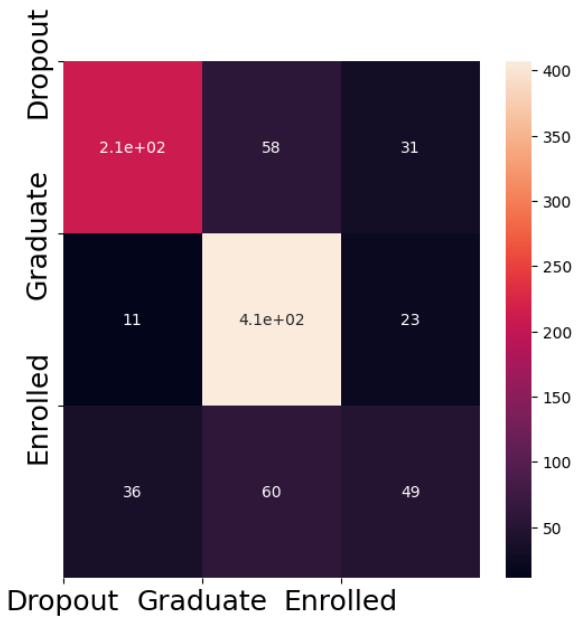

Figura 10: Matriz de confusión conjunto de testeo evaluado en modelo **RandomForestClassifier**

Red Neuronal:

Por último, implementamos una red neuronal con ayuda de la librería tensorflow.keras, la cual esta compuesta de 5 capas dendas de 64, 128, 64, 32 y 3 neuronas respectivamente y una función de activación relu en las primeras cuatro y una final de tipo sigmoid, para el entrenamiento es usado un optimizador tipo Adam, la función categorical_crossentropy como función de costo y accuracy como la correspondiente métrica, antes de entrenar se realizo nuevamente la división del conjunto de datos pero en tres subconjuntos, de entrenamiento, validación y testeo, donde la validación nos permite ir observando el comportamiento del modelo con los datos externos durante el proceso de entrenamiento, estos se dividieron en tamaños de 80%, 10% y 10% respectivamente. Posterior a esto con la ayuda de la funión StandardScaler de sklearn.preprocessing se estandarizaron las variables de entrada y además de esto se realizo una vectorización de los targets para poder realizar el entrenamiento.

El proceso de entrenamiento se realizo durante 50 épocas obteniendo un accuracy final para el conjunto de entrenamiento del 96.1% y para la evaluación del conjunto de testeo un accuracy de 92.54% muy superior a los resultados obtenidos por los modelos anteriores. En la Figura 11 se puede observar la curva de entrenamiento correspondiente observando la evolución del accuracy durante las épocas de entrenamiento. Por otro lado, en la Figura 12 se observa la matriz de confusión asociada, denotando una clara superioridad en la predicción de estos targets de este modelo respecto a los anteriores.

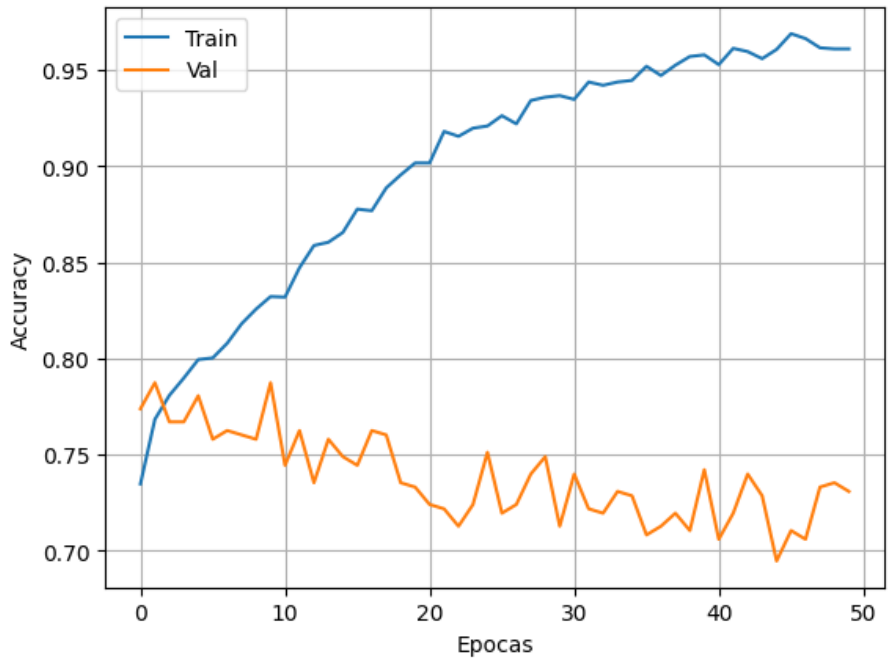

Figura 11: Curva de entrenamiento red neuronal

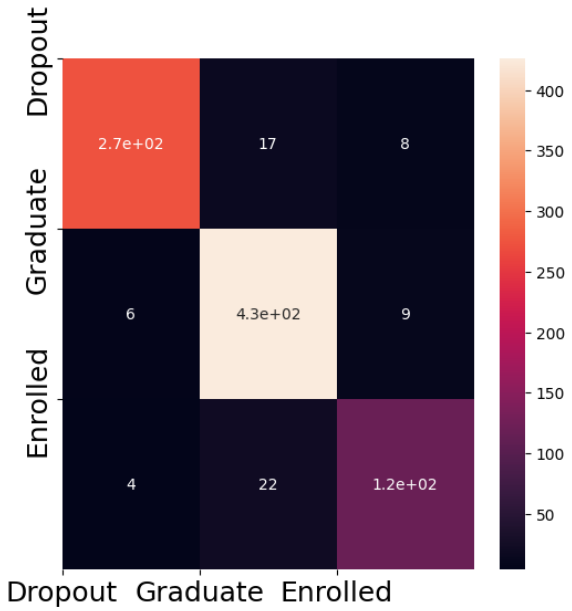

Figura 12: Matriz de confusión conjunto de testeo evaluado en modelo **Res Neuronal**

Conclusiones

- Se observa una fuerte correlación entre las variables correspondientes a Curricular de primer semestre y segundo semestre, en especial las correspondientes al mismo tipo de categoría cercana al 90%, siendo así solo necesario contar con las variables correspondientes a primer semestre para poder describir el problema.
- Se observa una disparidad entre la cantidad de valores de target por categoría siendo solo el 17.9% para Enrolled y del 49.9% para Graduate lo cual tuvo efecto sobre la predicción de los tres primeros modelos ensayados.
- Los modelos de tipo LogisticRegression, DecisionTreeClassifier, RandomForestClassifier aunque de forma general se comportan bien en la predicción de Graduate y Dropout, fallan en Enrolled debido a la baja cantidad de datos asociados a este en comparación.
- El modelo de red neuronal fue el mejor en la predicción de este target con un 92.54% de accuracy y con unas fallas mínimas observadas en la matriz de confusión para las 3 clases, permitiéndonos concluir que este modelo permite un entrenamiento optimo apezar de una baja cantidad de información para una variable respecto a las otras.